Evaluation of a Transplantation Algorithm for Expressive Speech Synthesis

Jaime Lorenzo-Trueba[1], Roberto Barra-Chicote[1], Junichi Yamagishi[2], Oliver Watts[2], and Juan M. Montero[1]

[1]Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain
[2]CSTR, University of Edinburgh, United Kingdom
jaime.lorenzo@die.upm.es

Abstract. When designing human-machine interfaces it is important to consider not only the bare bones functionality but also the ease of use and accessibility it provides. When talking about voice-based interfaces, it has been proven that imbuing expressiveness into the synthetic voices increases significantly its perceived naturalness, which in the end is very helpful when building user friendly interfaces. This paper proposes an adaptation based expressiveness transplantation system capable of copying the emotions of a source speaker into any desired target speaker with just a few minutes of read speech and without requiring the recording of additional expressive data. This system was evaluated through a perceptual test for 3 speakers showing up to an average of 52% emotion recognition rates relative to the natural voice recognition rates, while at the same time keeping good scores in similarity and naturality.

Keywords: expressive speech synthesis, emotions, adaptation, expressiveness transplantation

1 Introduction

In a world were technology is ever increasingly pervasive, the breach between impaired and healthy people also keeps increasing. For that reason if we want to promote social equality it is imperative that we try to bridge the aforementioned breach. In the field of speech synthesis the application is clear: building speech-based human-machine interfaces that allow all users to interact with computers with minimal training and difficulties [6],[11].

The present research proposes and evaluates an expressiveness transplantation technique that enables a speech synthesis system to produce expressive voices (i.e. imbuing emotions such as happiness or sadness, or speaking styles such as news broadcasting and political speech) that are much more adequate to the desired task. For example, if we were to produce an interface that interacts in a dialog system with the user, it would be much more natural if the system could emulate natural human interaction by including emotional cues into the

conversation, all in all increasing the naturalness of the system. This increase in naturalness in turn can translate into a reduction in the rejection of new technologies for the general public [14].

When looking at HMM-based speech synthesis, there has been a recent surge in research focusing on increasing naturality be it through expressiveness or through direct speech quality increases. One of the more trending approaches towards speech quality is using adaptation to obtain more robust systems capable of providing higher quality speech with less and less training data [19],[18]. The expressiveness approach, on the other hand, has shown good promise and is being researched from different points of view: modeling expressiveness as a feature included in the HMM decision trees [15], clustering training speakers acoustic features according to speaking styles [4], modification of glottal parameters to enhance noise robustness [16] or transplanting expressive information from preexisting sources into non expressive target voices [2]. All in all the objective is to obtain synthetic voices that completely mimic human speech and can be used unobtrusively in speech interfaces.

In section 2 we describe the speech databases used to train and test the proposed expressiveness transplantation algorithm, which is then described in section 3 together with the used adaptation technique. Section 4 thoroughly describes the design and environment of the perceptual evaluation carried on to test the presented technique. Finally in section 5 we analyze the obtained results and in section 6 we give a brief conclusion drawn from all the research process.

2 Speech Corpora

Both emotional and neutral speech corpora were used for the evaluation. The emotional data (SEV Database [1]) has been evaluated previously for the Albayzin2012 speech synthesis evaluation, making it ideal for the introduced evaluation. The neutral data on the other hand is a combination of published databases (UVIGO-ESDA Database [5]) and a pair of male speakers recorded in our laboratory environment.

SEV Database Emotional database consisting of a male and female speaker. Out of the available emotions only 4 of them are considered: anger, happiness, sadness and surprise also including the neutral voice as the reference. All the emotions were recorded for the same utterances favoring the learning of expressiveness cues, amounting approximately 30 minutes of training speech each.

UVIGO-ESDA Database A database consisting of a single male amateur Spanish speaker in a neutral situation totaling approximately 2 hours of speech recorded in studio.

Recorded Data A number of male and female speakers were recorded in our acoustically-treated room, providing high quality sources. The chosen speakers (JLC and JEC) both present significantly different fundamental frequencies and amounting a total of 30 and 7 minutes of speech respectively.

3 Transplanted Models Generation

There are two main objectives we want to fulfill with our system: obtain high quality expressive synthetic voices from low amounts of training data and secondly being able to transplant the expressiveness from previously available verified expressive models into target neutral speakers. Our approach to the first objective is using average models and adapting from them to increase the overall quality of the generated models, while the transplantation is rooted also in adaptation techniques and can be considered a generalization. Both points are discussed in this section.

3.1 Average Models Generation and Adaptation

One of the main advantages of parametrical speech synthesis such as the considered HMM-based synthesis is how easy it is to adjust the modeled parameters and with it change or adapt the models to new data. This means we can train a background average model including all the available data which can be expected to be very stable as there will be more training data than when just using a single source voice. If we then adapt this background model to some new speaker that has been recorded even with little amounts of speech, it follows that the obtained model will be much better than if we were to train it independently.

In our system we use a combination of Speaker Adaptive Training (SAT) [18] and Constrained Structural Maximum A Posteriori Linear Regression [19] as the background average training algorithm and adaptation algorithm respectively. The prior was chosen because SAT focuses on minimizing the adverse effects in the variance of the model that are introduced when using too heterogeneous training data. This is done by normalizing the influence of the speaker heterogeneity among the training speakers in both the state outputs and distributions. In the end the obtained average models are much more stable and provide a successful foundation for further adaptations.

For the adaptation technique we had a few requirements to consider: it must be capable of adapting from small amounts of data and it should also adapt variances in order to maintain the expressive nuances of the voices [19]. The second requirement forced us to use a constrained approach while the first one was not so direct to solve. The traditional MLLR [7], [12] approach is not good enough because it adapts the model as a whole and as such it is not optimal when the adaptation data is small, while MAP [9] only touches the contexts for which adaptation data is available. CSMAPLR is a combination between both techniques (with a MAP variation that allows the use of tree structures, SMAP [17]) that is specially suitable for using in combination with SAT and smaller amounts of adaptation data. Also, it is constrained so the variances are adapted fulfilling both of our requirements at the same time.

3.2 Transplantation Process

The original proposal of this paper is the expressiveness transplantation process, capable of learning the nuances in the differences between an expressive model

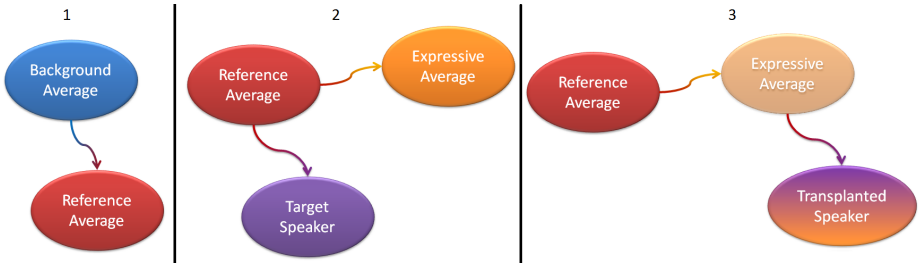

Fig. 1. Representation of the three steps in the transplantation process. The ovals represent HTS models and the arrows represent CSMAPLR adaptation transforms.

and a reference model and then transplanting them into a different target speaker reference model (see figure 1). Originally this was implemented directly at the synthetic voice generation step [2], but we have evolved this concept in order to include the transplantation into the training process of the models, attaining significantly better performances.

We consider that the adaptation transformation function that relates the reference and the expressive model is capable of capturing the expressive cues to be transplanted, but in order to obtain a generalizable and streamlined expressiveness transplantation there are a few points to take into account. First of all, if we want to apply the same transformation function between different models, and because the adaptation process (CSMAPLR) relies on the decision tree structures, the trees must be shared between the models. This is easily solved by adapting all the required models from the same background model. Also it is important to see that it is likely that different people would speak differently even under the same expressive category. As such it is interesting to create expressive averages that include multiple speakers so that the transformation function that relates it to the neutral/reference average is more robust. Finally, an advantage of using this transformation functions is that they can be obtained offline for the different expressive categories resulting in a very fast transplantation process once they are available.

For the transplantation system itself the process is as follows:

1. Adapt the reference average from the background average (Fig. 1.1).
2. Adapt the target speaker and the different expressive averages from this reference model. These are the transformations that must be kept for the transplantation step (Fig. 1.2).
3. Apply the transformation function between the expressive average and the reference with the desired transplantation ratio (K) and then to the resulting model apply the transformation function between the reference and the target speaker. The obtained model is the expressive target speaker (Fig. 1.3).

The inclusion of the transplantation ratio is mainly due to the smoothing in parametrical modeling. This smoothing causes a reduction in the expressive

strength of the synthetic voices that can be perceived even in non-transplanted synthetic voices as seen for example in Albayzin2012 expressive speech synthesis challenge [10]. For that evaluation we proposed a transplantation ratio that allowed us to either enhance or reduce the expressive strength of the models [13], and the same concept can be applied when transplanting: scaling the transformation function of the expressive step enables us to control the expressive strength of the transplanted voice. The models obtained this way can be then used to normally synthesize expressive utterances while reasonably keeping the identity of the target speaker. The validity of the expressiveness recognition, emotional strength and similarity transplantation hypothesis are validated in the results of the perceptual test in section 5.

4 Perceptual Test Description

The goal of the test was to verify if the expressiveness was transplanted successfully in terms of speaker similarity, expressiveness recognition rates and expressive strength together with a naturality test. This was done in an environment in which the expressiveness were the emotions (anger, happiness, sadness, surprise and the neutral as a reference) of joaquin in the the SEV emotional database 2 and the target speakers were three neutral speakers (UVD, JEC and JLC). The test itself was designed following a balanced latin-square [8] approach to the following systems:

- The natural voices of all the speakers (joaquin and his 4 emotions, UVD, JEC, JLC).
- The synthetic neutral voice of the speakers.
- The 4 transplanted emotions into the different speakers with transplantation ratios (K) of 0.50, 0.75, 1.00 and 1.25.

Every target speaker was analyzed separately, so for each testing session the total number of systems was 24. Following the latin-square approach this meant that we needed 24 different utterances to be synthesized (or selected from the natural database) for all the systems, to be presented to the listeners in a completely random order without repeating either system or utterance throughout the test. Consequently the minimum number of listener for each test was 24 and there were 3 testing waves.

The test itself was done through a web interface presenting the listener either one utterance or several depending on the section of the test. These utterances could be played as many times as desired by the listener so that they could answer the questions with confidence despite the difficulty of the task. The first section asked questions about the naturality, emotional strength and recognized emotion by reproducing a single utterance picked from the previously defined 24. Naturality and emotional strength were ranked in a 5 point likert scale and the recognized emotion could be selected from a list of the following: anger, happiness, sadness, surprise, neutral, other. The second part of the test focused on

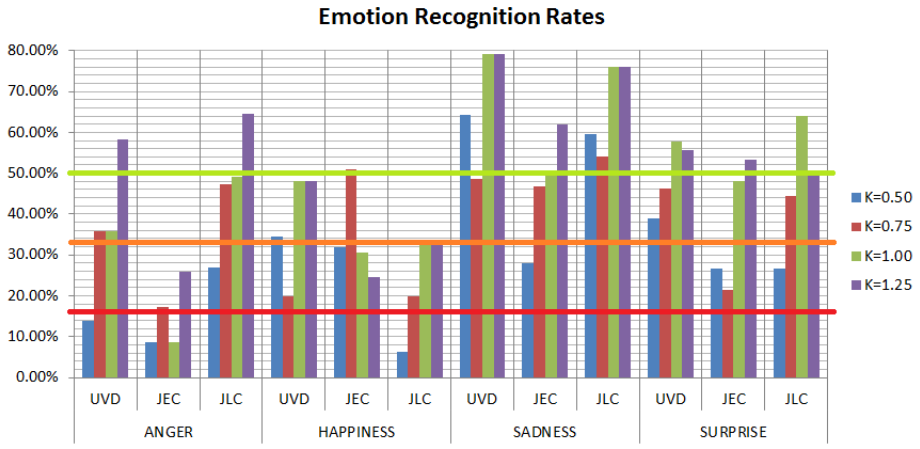

Fig. 2. Emotion recognition rates for the 3 transplanted speakers relative to the identification rates of the natural voice for different transplantation ratios (K). The red horizontal line represents the random threshold, the orange line twice the threshold and the green line three times that value.

similarity and placed the listener against a synthetic test sample and 4 reference natural voice samples of the target speaker that had then to be ranked in similarity once again in a 5 point likert scale.

5 Results

Both table 1 and figure 2 show the results for the 4 evaluated aspects: emotion identification rates (EIR), emotional strength (ES), speech quality (SQ) and speaker similarity (SIM). We only present the results broken down for all the speakers for the EIR (figure 2) as for the rest of the parameters they presented similar trends without particular significance. As such the results in table 1 show the average between the three speakers.

By taking a look at figure 2 we can see that the recognition rates are in the vast majority of the cases higher than the random threshold, being most of the time more than twice that value and sometimes higher than 50% (always relative to the natural voice EIR).

– Even if these values may not seem high at a glance, it is important to notice that the task at hand is extremely hard not only for us in the transplantation process but also for the listeners when recognizing emotions. Our results, when compared to non-transplanted synthetic data (Synthethic rows in table 1) prove that the proposed system is being very successful at creating emotions from pure neutral data.

EMOTION	Transplantation Ratio	Identification Rate	Speech Quality	Emotional Strength	Speaker Similarity	Average Performance
Anger	0,50	18,6	70,4	63,4	93,5	61,5
	0,75	38,4	68,4	63,4	90,3	65,1
	1,00	36,0	67,3	65,2	84,6	63,3
	1,25	58,1	62,2	67,6	75,3	65,8
	Synthetic (Src)	90,7	77,6	83,9	86,4	84,6
	Natural (Src)	100	100	100	100	100
Happiness	0,50	33,8	73,2	69,9	109,7	71,7
	0,75	42,3	70,7	71,9	103,8	72,2
	1,00	52,1	62,8	69,1	94,5	69,6
	1,25	49,3	60,8	75	89,0	68,5
	Synthetic (Src)	91,5	74,6	87,5	120,7	93,6
	Natural (Src)	100	100	100	100	100
Sadness	0,50	65,4	72,5	65,0	105,3	77,1
	0,75	64,1	67,8	65,8	98,0	73,9
	1,00	88,5	60,4	67,7	86,2	75,7
	1,25	92,3	70,1	69,7	76,1	77,1
	Synthetic (Src)	83,3	70,1	74,8	104,0	83,1
	Natural (Src)	100	100	100	100	100
Surprise	0,50	39,7	69,2	62,6	70,5	60,5
	0,75	47,4	62,6	68,0	69,6	61,9
	1,00	73,1	62,8	66,4	67,8	67,5
	1,25	67,9	56,1	70,4	65,5	65,0
	Synthetic (Src)	64,1	77,4	76,9	73,1	72,9
	Natural (Src)	100	100	100	100	100

Table 1. Results (in %) normalized against the natural voice results. The results of the natural and synthetic source (labeled as Src in the table) voices are included for comparison purposes and have been extracted from the Albayzin2012 results [13]. Average performance is the average of the 4 evaluations.

- Further significant trends are that the performance for the prosodical emotions (surprise, sadness) is much higher than for the spectral emotions (happiness, anger) mainly because it is easier to distort the voice when manipulating spectral features than when manipulating prosodical features.
- Also it is important to notice that the overall recognition rates seem to have a degree of correlation with the amount of training data available, being significantly better for the speaker with more data (UVD at a 60% average recognition rate, compared to a 41% average for JEC).
- We can also see a trend that will be reinforced with the rest of the analyzed parameters, and that is that the expressiveness does increase with the transplantation ratio (K), in this case increasing the identifiability of the emotions.

By looking at table 1 we can see the results for SQ (4th column), ES (5th column) and SIM (6th column) with an averaged performance to give a hint

on the overall performance of tnat particular transplantation ratio and emotion combination:

- If we compare the obtained results to those of the source natural expressive voices (Natural rows), we can see how there is an average drop of approximately 30% in SQ and ES. The decrease in SQ is can be assumed to be due to the HMM-based synthesis process. The decrease in ES is expected to be caused by the inherent smoothing of the parametrization and adaptation steps, and it can be somewhat compensated by using higher transplantation ratio values.
- Comparing with the synthetic expressive voices (Synthetic rows), on the other hand, shows only about 10% decreases in average. This means that the transplantation process is capable of imbuing the neutral target voice with expressiveness without reducing speech quality significantly further than what traditional HMM-based synthesis provides.
- Similarity results are very good when comparing to either source system, in some situations providing even better speaker identifiability than natural voices. While this is due to the difficulty of identifying the speaker in an expressive environment, the proposed system clearly excels in this evaluation factor.

To sum up, the results have shown that both the transplantation of expressiveness and the transplantation strength control are successful, with the following particularities: for increasing transplantation ratios both the emotion identification rates and emotional strength increase, particularly for K's greater or equal than 1,00. Conversely, for decreasing transplantation ratios it is the similarity and the speech quality that get better results. In the end this means that while we can define a global optimum that maximizes all the defined evaluation parameters between K=0.75 and K=1.00 that would work for most target voices, it is definitely more interesting to use a transplantation ratio that enhances the particular feature more adequate to the intended task.

6 Conclusions

We have proposed an expressiveness transplantation process capable of learning the differences between a particular expressive speaking style and a reference and applying it to a new target speaker, producing expressive synthetic voices that have the identity of the target speaker. This was done by making use of adaptation techniques and applying them in cascade, which does not harm the produced speech quality. This system was applied to an emotional database and a set of 3 neutral speakers and tested by means of a perceptual test that proved that the emotions are successfully transplanted.

The results show that the process produces voices of the same quality and similarity as traditional HMM-based speech synthesis (average performances about 10% lower in the evaluated features) while providing emotion identification rates topping at 60% for one of the speakers.

The transplantation ratio that was used to control the expressive strength of the transplantation was also proven to provide the desired results, as increasing its value provided higher identification rates and strength scores at the cost of similarity and quality due to the higher degree of manipulation of the models.

In conclusion we have introduced a system capable of imbuing expressiveness to any target voice, enabling us to produce systems much more real-life sounding in their intended end-user applications. A web-based demo is available online [3].

Planned future work includes increasing the versatility of the transplantation process allowing for the selection of feature streams to be adapted (i.e. treating prosodical and spectral features differently), and designing a user feedback interface and process to effortlessly ascertain the optimal transplantation ratio and adaptation parameters for the desired task and target voice that do not require any technical knowledge for the end-user. We also intend to test the transplantation in a controlled speaking styles environment and to define a preference test environment to check not only the validity of the system but also how it fares when compared to more traditional approaches.

7 Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo-Trueba has been funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

References

1. Barra-Chicote, R., Montero, J.M., Macias-Guarasa, J., Lufti, S., Lucas, J.M., Fernandez, F., D'haro, L.F., San-Segundo, R., Ferreiros, J., Cordoba, R., Pardo, J.M.: Spanish expressive voices: Corpus for emotion research in spanish. Proc. of LREC (2008)
2. Barra-Chicote, R.: Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis. Ph.D. thesis, ETSIT-UPM (2011)
3. Barra-Chicote, R., Lorenzo-Trueba, J., Montero, J.M.: Acoustic emotional patterns transplantation to new speakers (2013), `http://lorien.die.upm.es/barra/emo-transplantation/index.php`
4. Chen, L., Gales, M., Wan, V., Latorre, J., Akamine, M.: Exploring rich expressive information from audiobook data using cluster adaptive training. In: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13 (2012)
5. E.T. Banga, C.M.: Documentation of the uvigo-esda spanish database. Tech. rep., Grupo de Tecnoloxias Multimedia, Universidad de Vigo, Vigo, Espaa (2010)

6. Forgrave, K.E.: Assistive technology: Empowering students with learning disabilities. The Clearing House 75(3), 122–126 (2002)
7. Gales, M., of Cambridge. Engineering Dept, U.: Maximum likelihood linear transformations for hmm-based speech recognition. Computer speech and language 12(2) (1998)
8. Gao, L.: Latin Squares in Experimental Design. Michigan State University (2005)
9. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. Speech and Audio Processing, IEEE Transactions on 2(2), 291–298 (1994)
10. Iberspeech2012: 2012 albayzin evaluations: speech-synthesis results (2012)
11. Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M.: A speech synthesis system with emotion for assisting communication. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)
12. Leggetter, C., Woodland, P.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer speech and language 9(2), 171 (1995)
13. Lorenzo-Trueba, J., Watts, O., Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M.: Simple4all proposals for the albayzin evaluations in speech synthesis. In: Iberspeech2012, VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop (2012)
14. Lutfi, S.B.L.: User-centric Need-driven Affect Modeling for Spoken Conversational Agents: Design and Evaluation. Ph.D. thesis, ETSIT-UPM (2013)
15. Nose, T., Kobayashi, T.: An intuitive style control technique in hmm-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. Speech Communication (2012)
16. Raitio, T., Suni, A., Vainio, M., Alku, P.: Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. Computer Speech & Language (2013)
17. Shinoda, K., Lee, C.H.: A structural bayes approach to speaker adaptation. Speech and Audio Processing, IEEE Transactions on 9(3), 276–287 (2001)
18. Yamagishi, J., Kobayashi, T.: Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training. IEICE TRANSACTIONS on Information and Systems 90(2), 533–543 (2007)
19. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. Audio, Speech, and Language Processing, IEEE Transactions on 17(1), 66–83 (2009)